# Differentially Private k-Means with Constant Multiplicative Error

Haim Kaplan (Tel Aviv University and Google)    Uri Stemmer (Ben-Gurion University)

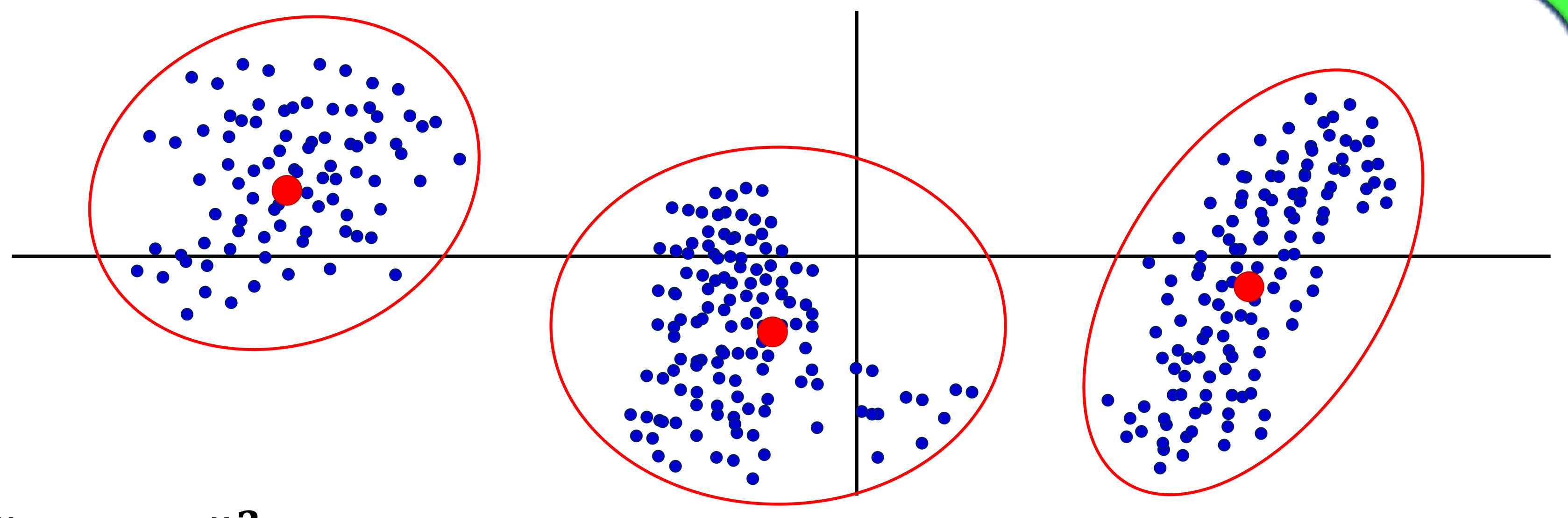
## What is differential privacy?

[DMNS06]

A (rand) algorithm $\mathcal{A}$ is $(\epsilon, \delta)$ differentially private if for all neighboring databases $S_1, S_2$ and for all sets of outputs $F$:

$$\Pr[\mathcal{A}(S_1) \in F] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(S_2) \in F] + \delta$$

## What is k-means clustering?



**Given:** Data points $S = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$

**"Task":** Identify groups of data points, and assign each point to one of the groups

**Intuition:** Clusters have "centers", and points are nearer to the center of their cluster

**Goal:** Identify $k$ centers $C = (u_1, \dots, u_k) \in (\mathbb{R}^d)^k$ that minimize $\text{cost}(C) = \sum_{i \in [n]} \min_{\ell \in [k]} \|x_i - u_\ell\|^2$

| Ref | Runtime | Bounds (informal) |
|---|---|---|
| MT'07 | $n^{kd}$ | $\text{OPT} + \tilde{O}(k \cdot d)$ |
| GLMRT'10 | $n^d$ | $O(1) \cdot \text{OPT} + \tilde{O}(k^2 \cdot d)$ |
| BDLMZ'17 | poly | $O(\log^3 n) \cdot \text{OPT} + \tilde{O}(k^2 + d)$ |
| FXZR'17 | poly | $O(k) \cdot \text{OPT} + \tilde{O}\left(k^{3/2} \cdot \sqrt{d}\right)$ |
| New | poly | $O(1) \cdot \text{OPT} + \tilde{O}\left(k \cdot \sqrt{d}\right)$ |

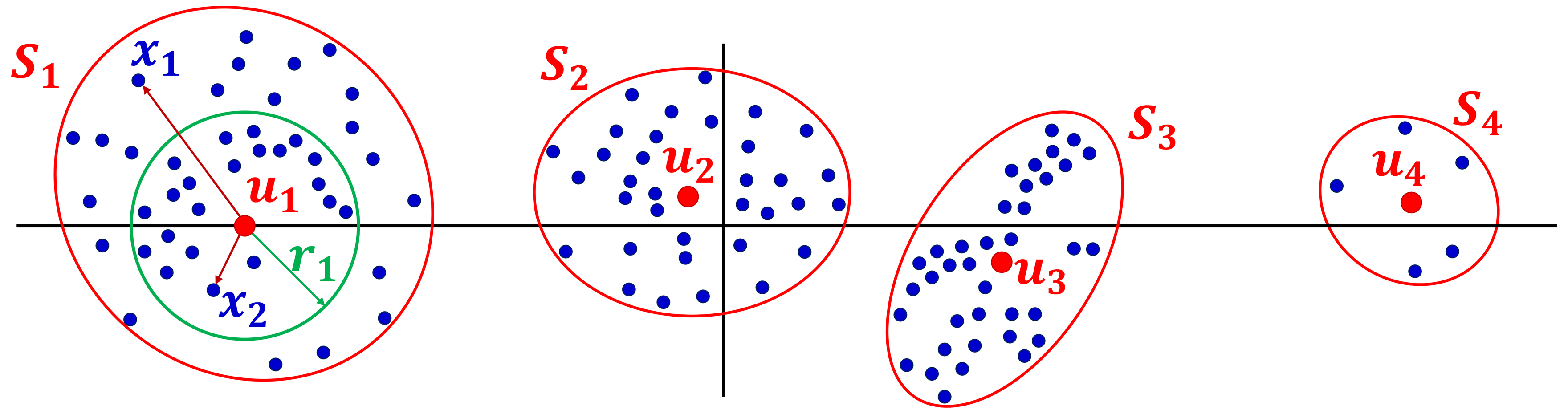
## Previous work: local search [GLMRT'10], [BDLMZ'17]

1. Let $Y$ be a finite discretization of the unit ball
2. Let $C \subseteq Y$ be an arbitrary set of $k$ centers
3. For $T \approx k \cdot \log n$ rounds:
   a) Choose $(x, y) \in C \times Y$ approximately minimizing $\text{cost}(C \backslash \{x\} \cup \{y\})$.
   b) Set $C \leftarrow C \backslash \{x\} \cup \{y\}$

**Result:** Constant multiplicative error w.r.t. centers in $Y$. Runtime $\approx |Y|$

$\Longrightarrow$ Suffices to privately identify a small set of candidate centers $Y$ containing a subset of k candidates with low k-means cost

## Find $Y$ containing $k$ centers with low cost



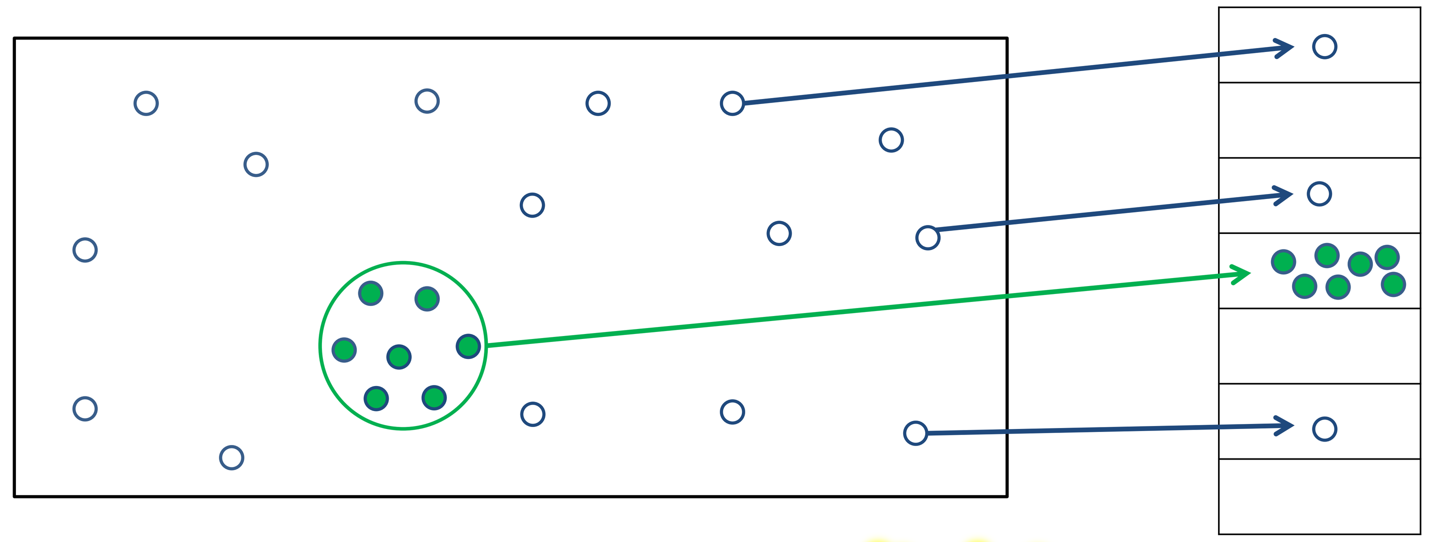

- Let $u_1, \dots, u_k$ denote $k$ optimal centers, and $S_1, \dots, S_k$ the induced clusters
- **Obs1:** Can ignore small clusters (pay on additive error)
- **Obs2:** Let $r_i$ = min s.t. $|\mathfrak{B}(u_i, r_i) \cap S_i| \geq |S_i|/2$. Only need $y_i \in Y$ s.t. $\|y_i - u_i\| \leq O(r_i)$

**Suffices to solve:** Privately identify a small set $Y \subseteq \mathbb{R}^d$ such that: For every "large enough" cluster $P \subseteq S$, w.h.p. $\exists y \in Y$ s.t. $\|y - \text{avg}(P)\| \leq O(\text{diam}(P))$

- **Obs3:** Suffices to capture every "large enough" cluster $P$ of diameter (roughly) $r$, and to execute in parallel with exponentially growing choices for $r$

## Useful Tool: LSH [Indyk&Motwani]

- **Maximize** the probability of **collision** for **similar** items
- **Minimize** the probability of **collision** for **dissimilar** items



**Hopefully:** "Heavy" buckets correspond to clusters

## Additional Results

- k-means under local differential privacy with constant multiplicative error
- Results also hold for k-medians
- Private coresets for k-means and k-medians

## Some of the Challenges

- How to capture small clusters?
- How to implement local search in the local model?

## Acknowledgements

- Research partially supported by the Frankel Center for Computer Science